Exploiting a geometrically sampled grid in the steered response power algorithm for localization improvement

D. Salvati,[a] C. Drioli, and G. L. Foresti
Department of Mathematics, Computer Science and Physics, University of Udine, Udine, Italy

(Received 1 January 2016; revised 16 November 2016; accepted 6 January 2017; published online 26 January 2017)

The steered response power phase transform (SRP-PHAT) is a beamformer method very attractive in acoustic localization applications due to its robustness in reverberant environments. This paper presents a spatial grid design procedure, called the geometrically sampled grid (GSG), which aims at computing the spatial grid by taking into account the discrete sampling of time difference of arrival (TDOA) functions and the desired spatial resolution. A SRP-PHAT localization algorithm based on the GSG method is also introduced. The proposed method exploits the intersections of the discrete hyperboloids representing the TDOA information domain of the sensor array, and projects the whole TDOA information on the space search grid. The GSG method thus allows one to design the sampled spatial grid which represents the best search grid for a given sensor array, it allows one to perform a sensitivity analysis of the array and to characterize its spatial localization accuracy, and it may assist the system designer in the reconfiguration of the array. Experimental results using both simulated data and real recordings show that the localization accuracy is substantially improved both for high and for low spatial resolution, and that it is closely related to the proposed power response sensitivity measure. © *2017 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4974289]

[JL] Pages: 586–601

I. INTRODUCTION

The problem of locating acoustic sources is a fundamental task in applications of acoustic scene analysis and acoustic situational awareness, and it received significant attention in the audio processing research community. Acoustic source localization using a microphone array can be performed by indirect and direct methods. The indirect (two-step) approach computes a set of time difference of arrival estimates (TDOAs) using measurements across various combinations of microphones, and then estimates the source position using geometric considerations.[1–3] Direct methods are based on maximizing the steered response power (SRP) of a beamformer and they are very attractive in acoustic applications due to their robustness in noisy and reverberant conditions.[4–9]

In this work, we consider the localization of a single source in a reverberant environment. This scenario can be of interest in different practical applications such as videoconferencing systems, in which the estimation of sound coordinates can be used to automatically steer a video camera towards an active speaker, or in human-computer interaction systems, in which the localization is used in beamforming based signal enhancement for speech recognition or dictation system. The SRP phase transform[5] (SRP-PHAT) is one of the most effective direct methods for the localization of an acoustic source in reverberant environments. It is based on a steered beamformer, which can be implemented using a space search procedure, and a map that links each position of the search grid to the TDOA functions related to the sensor pairs. The use of an acoustic map related to the TDOA between two microphones was first introduced in 1998 by Omologo *et al.*[4] The authors called this procedure global coherence field (GCF).[11] In 2001, DiBiase *et al.*[5] demonstrated that the SRP-PHAT can be computed by using the GCF and the generalized cross-correlation phase transform (GCC-PHAT),[10] making its practical implementation very attractive. In fact, the GCC-PHAT can be computed in the frequency domain using the discrete Fourier transform (DFT) for each sensor pair, and the acoustic map can be computed by memory accesses and scalar additions on a look-up table storing the GCC-PHAT values. The sampled space grid, which is a set of candidate positions for the source, is pre-calculated defining a look-up table that links the position in space with TDOA values of microphone pairs. The role of the PHAT filter is to normalize the narrowband steered beamformer and to only take into account the phases of the cross-power spectral density. The normalization has the positive effect of increasing the spatial resolution for broadband sources,[12] when the source signal is self-correlated and the self-correlation time is larger than a given threshold (e.g., for voiced sounds). Hence, the normalization can help the localization in a reverberant environment since it allows improved identification of direct paths and reflections.

Most of the research on SRP-PHAT focused on solutions to reduce the computational cost of the grid-search step.[13–15] However, these methods usually discard part of the information available and the localization performance can degrade when reverberation increases.[16] Recently, a method that relies on the use of a coarser grid has been proposed in Ref. 17. Herein it is shown that the traditional grid-

[a]Electronic mail: daniele.salvati@uniud.it

search approach of SRP-PHAT degrades its performance when the spatial resolution decreases due to the loss of information of GCC-PHAT functions. To face this problem, in Ref. 17 a modified SRP (called M-SRP) is proposed to accumulate the GCC-PHAT values in the volume surrounding each point of the defined spatial grid. Reducing the spatial grid leads to a lower computational cost, but also reduces the accuracy, which is limited by the resolution of the grid. Other methods have been proposed that improve the localization accuracy by refining the search procedure from a coarser grid to a finer grid using iterative searching procedures.[16,18,19]

The abovementioned methods have in common the way in which the space search grid is designed, and the way in which the relationship between the points on the grid and the TDOAs of microphone pairs is built. Specifically, for each microphone pair and for each point on the grid, an unique integer TDOA value is selected to be the acoustic delay information linked to that point. This uniform regular grid (URG) procedure does not guarantee that all TDOA samples are associated to points on the grid, nor that the spatial grid is consistent since some of the points in the grid may not correspond to an intersection of a bare minimum of three hyperboloids (or two hyperbolas, in 2D). The linking from space points on the grid to TDOAs also does not allow for spatial resolution scalability, since when the number of points is reduced, part of the TDOA information gets lost as it results no more associated to any points on the grid. For these reasons, different methods have been proposed in Refs. 16–18 to collect and use the TDOA information related to the volume surrounding each spatial point on the search grid. A boundary-vertex approach is used in Ref. 16 (called H-SRP), in which the GCC-PHAT accumulation limits are determined by the cube surrounding the volume vertices. In Ref. 18, a similar approach of M-SRP is proposed that exploits the mean of the accumulated GGC-PHAT values for each volume (we refer to it here as I-SRP). However, these methods do not take into account how TDOA information is distributed in the space. We will see that the spatial distribution of all TDOA information is important knowledge that can be used to compute a sensitivity measure of the acoustic system with respect to the search region and to improve the localization accuracy. There is thus the need for a rigorous analysis of the spatial grid map and of how the TDOA information from GCC-PHAT functions is accumulated in the space.

In this paper, we propose a new spatial grid design procedure in the SRP-PHAT, named geometrically sampled grid (GSG), which makes use of the discrete hyperboloids (representing all possible locations related to a TDOA) and of their intersections, to design an acoustically coherent space grid on which the source search can be performed. The GSG method builds the steered power response function using all the TDOA information available from the GCC-PHAT functions related to the sensor pairs in the array. Moreover, we will show how, based on the density analysis of hyperboloid intersections, a steered power response sensitivity analysis of the localization system can be conducted. We refer to "sensitivity" as a quantified measure of the change of the response power with respect to the change of the spatial position, predicting where the search space will be characterized by higher and lower localization accuracy. To date, studies concerning the information distribution of SRP-like localization methods are not frequent in the literature. An example is Ref. 20, in which a discriminability measure is proposed, which only considers the array geometry and the sampling frequency to distinguish a given point in space from its neighbors. In contrast with it, the proposed GSG includes in the analysis process a relationship between the sampled space and all discrete samples of the GCC-PHAT functions to prevent the loss of information that may arise from the choice of an arbitrary desired spatial resolution.

Besides that, the coherent sample grid and the power response sensitivity analysis are useful tools to decide if the spatial resolution and the sensitivity map of a given array configuration are adequate and, if not, to assist the system designer in its reconfiguration (e.g., by the positioning of additional sensors or by increasing the sampling frequency). Hence, it means that the system configuration designed by the GSG procedure generates a grid in which each point is consistent for the localization, i.e., it is the point of intersection of at least three hyperboloids.

The use of all the TDOA information available from the GCC-PHAT functions solves the problem of arbitrarily selecting the spatial grid resolution without loss of information, and it turns out to notably improve the localization performances. The geometric approach based on the analysis of hyperboloid intersections allows the design of a sensitivity map, in which the regions where the localization is more accurate correspond to the high sensitivity regions of the steered power response function. Moreover, the sensitivity map is also a useful tool for indirect methods since they are naturally based on discrete sampling of TDOA to compute the source position estimation. However, indirect methods based on GCC-PHAT only take into account the maximum value information of the GCC-PHAT function, and the localization performance considerably degrades in noisy and reverberant conditions.[5] On the other hand, the SRP-PHAT based on GSG builds an improved acoustic map using the whole GCC-PHAT information available after TDOA discretization, and estimates the source position by searching the maximum value of this acoustic map. This leads to an increment of robustness in adverse conditions.

Finally, the GSG method might also provide reduced computational cost with respect to the URG method in three cases: (1) when the search procedure is restricted to the coherent grid, thus discarding the URG points which are not covered by sufficient acoustic information, (2) when the type of application allows one to use a coarser grid and a lower spatial resolution, (3) when the search can be restricted only to the high sensitivity regions, in which the localization accuracy is maximized.

The paper is organized as follows. After presenting the relationship between the spatial grid and the TDOA functions in Sec. II, the GSG algorithm is described in Sec. III. In Sec. IV, the GSG based SRP-PHAT is presented. Finally,

Sec. V illustrates experimental results obtained in a simulated reverberant environment and in a real-world scenario.

II. SPATIAL GRID AND TIME DIFFERENCE OF ARRIVAL

Consider a reverberant room, and a volume $G = G_x \times G_y \times G_z$, discretized with a space resolution Δ, in which the acoustic source is being searched. A generic grid position is denoted by $\mathbf{r}_g = [x_g, y_g, z_g]^T$, $\mathbf{r}_g \in G$, where $(\cdot)^T$ denotes the transpose operator. Within the room, we suppose M microphones are disposed according to a given geometry. The positions of the M microphones in Cartesian coordinates are $\mathbf{r}_m = [x_m, y_m, z_m]^T$, $m = 1, 2, \ldots, M$. We will consider all possible sensor pairs of the array in our analysis. Accordingly, an array of M microphones provides N unique microphone pairs, with

$$N = \binom{M}{2}. \tag{1}$$

Given a generic sensor pair n, referring to two microphones located in $\mathbf{r}_i$ and $\mathbf{r}_j$, the maximum TDOA in samples $T_n \in \mathbb{Z}$ is obtained as

$$T_n = \left\lceil \frac{\|\mathbf{r}_i - \mathbf{r}_j\| f_s}{c} \right\rceil, \tag{2}$$

where $\lceil \cdot \rceil$ denotes the floor function that maps a real number to the largest previous integer, f_s is the sampling frequency, c is the speed of sound, and $\| \cdot \|$ denotes Euclidean norm. The admissible range of values for the TDOA is $[-T_n, T_n]$, thus the possible discrete TDOA values for the sensor pair n are $2T_n + 1$.

We study the case in which a single acoustic source is active at time k and the unknown coordinate position is $\mathbf{r}_s(k) = [x_s(k), y_s(k), z_s(k)]^T$. The observed signals are given by the convolution of the unknown source $s(k)$ with corresponding acoustic impulse responses h_m from the source to the microphone m. We consider a linear and time-invariant system. The single-source reverberant model for discrete-time signals can be expressed as

$$\tilde{x}_m(k) = h_m * s(k) + v_m(k), \tag{3}$$

where $m = 1, 2, \ldots, M$, $*$ denotes convolution, and $v_m(k)$ is an additive noise term, uncorrelated with the source signal $s(k)$. Due to the propagation time of the source from its position to sensor position (expressed by the direct-path in the acoustic impulse response h_m), the wavefront reaches two microphones at different times. The difference τ of such instants is, in principle, related to the time difference between the largest peaks in the impulse responses h_m, corresponding to the direct paths of propagation. The relationship between a generic space position $\mathbf{r}_g$ and the discrete TDOA of the wavefront at the sensor pair n of two microphones i and j can be expressed as

$$\tau_n(\mathbf{r}_g) = \left\lceil \frac{(||\mathbf{r}_g - \mathbf{r}_i|| - ||\mathbf{r}_g - \mathbf{r}_j||) f_s}{c} \right\rfloor, \tag{4}$$

where $\lceil \cdot \rfloor$ denotes the rounding to the nearest integer. From Eq. (4), we can see that the locus of possible sound source locations generating the same TDOA for that microphone pair is described by a half-hyperboloid.

The spatial grid in the SRP-PHAT algorithm is traditionally calculated with an URG approach that links the uniformly distributed points on the spatial grid to TDOAs related to the sensor pairs using Eq. (4). The limitations of this approach are that it does not guarantee that all TDOA values correspond to a point on the space grid (and if this is the case, the information related to that TDOA is lost), and that it is not guaranteed that every point of the grid is consistent with the condition of being the locus where at least three half-hyperboloids intersect. Note that, due to the rounding operator, from the URG point of view everything goes as if in each grid position there is an intersection of N hyperboloids. The approximation due to the rounding operation can link a whole set of neighbor points to the same TDOA, resulting in practice in an uniform steered response power in that region.

III. GEOMETRICALLY SAMPLED GRID ALGORITHM

The geometrically sampled grid (GSG) algorithm is based on computing the space grid map by using the discretization of hyperboloids with a desired spatial resolution, and by taking all discrete TDOA values into account.

Consider a generic microphone pair n. We can interpret Eq. (4) as the quadratic surface of a hyperboloid in a local Cartesian system (x_n, y_n, z_n) with the origin in the midpoint of the segment joining the two microphones i and j,

$$\frac{x_n^2}{a_1^2} - \frac{y_n^2}{a_2^2} - \frac{z_n^2}{a_3^2} - 1 = 0, \tag{5}$$

where $a_1 > 0$, $a_2 > 0$, and $a_3 > 0$. This is the equation of a hyperboloid of two sheets in which the x_n axis is coincident with the line joining the two microphones. The transformation between the two coordinate systems (x, y, z) and (x_n, y_n, z_n) can be expressed as the combination of a translation and a rotation, i.e.,

$$\begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix} = \Omega_n \mathbf{R}_n \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \tag{6}$$

where Ω_n and $\mathbf{R}_n$ are, respectively, the translation matrix and the rotation matrix for pair n. Equation (5) can be rewritten in a simpler form as a hyperbola rotated about the x_n axis. In such case, we have a rotational hyperboloid and $a_3 = a_2$. By including the information in $\tau_n \in [-T_n, T_n]$ for the sheet identification, the hyperbola on axes (x_n, y_n) can be written in the following way:

$$x_n = f_x^{\tau_n}(y_n) = \frac{\tau_n}{|\tau_n|} \sqrt{\left(\frac{y_n^2}{a_2^2} + 1\right) a_1^2}. \tag{7}$$

Comparing Eq. (4) (at $z = 0$) and Eq. (7) we have

$$a_1 = \frac{c\tau_n}{2f_s},$$
$$a_2 = \sqrt{\left(\frac{\|\mathbf{r}_i - \mathbf{r}_j\|}{2}\right)^2 - a_1^2}. \tag{8}$$

We call $i_y\Delta$, $i_y \in \mathbb{Z}$, the discretization of y_n with resolution step Δ, and we calculate the grid points x_n from Eq. (7) with resolution Δ as

$$x'_n = \left\lceil \frac{f_x^{\tau_n}(i_y\Delta)}{\Delta} \right\rfloor \Delta. \tag{9}$$

We can now refer to the circumference with radius $i_y\Delta$ to obtain the rotation of the hyperbola along the x_n axes. If $i_z\Delta$ is the discretization of z_n, $i_z \in \mathbb{Z}$, we have

$$y'_n = \pm \left\lceil \frac{\sqrt{(i_y\Delta)^2 - (i_z\Delta)^2}}{\Delta} \right\rfloor \Delta. \tag{10}$$

The discrete half-hyperboloid Λ'_{n,τ_n} is thus given by

$$\Lambda'_{n,\tau_n} = \left\{ (x'_n, y'_n, z'_n) \in \mathbb{R}^3 : \left(x'_n = \left\lceil \frac{f_x^{\tau_n}(i_y\Delta)}{\Delta} \right\rfloor \Delta, \right.\right.$$
$$y'_n = \pm \left\lceil \frac{\sqrt{(i_y\Delta)^2 - (i_z\Delta)^2}}{\Delta} \right\rfloor \Delta \Bigg),$$
$$\left. z'_n = i_z\Delta, i_y \in \mathbb{Z}, i_z \in \mathbb{Z} \right\}. \tag{11}$$

With this procedure the Δ spatial resolution is guaranteed for the y-axis and the z-axis, but not for the x axis. We can then rewrite Eq. (7) in the following form:

$$y_n = f_y^{\tau_n}(x_n) = \frac{\tau_n}{|\tau_n|} \sqrt{\left(\frac{x_n^2}{a_1^2} - 1\right) a_2^2}. \tag{12}$$

We now call $i_x\Delta$, $i_x \in \mathbb{Z}$, and $i_z\Delta$, $i_z \in \mathbb{Z}$, the discretizations of x_n and z_n, respectively. We can compute the discrete half-hyperboloid Λ''_{n,τ_n} as

$$\Lambda''_{n,\tau_n} = \left\{ (x''_n, y''_n, z''_n) \in \mathbb{R}^3 : \left(x''_n = i_x\Delta, \right.\right.$$
$$y''_n = \pm \left\lceil \frac{\sqrt{(f_y^{\tau_n}(i_x\Delta))^2 - (i_z\Delta)^2}}{\Delta} \right\rfloor \Delta,$$
$$\left. \left. z''_n = i_z\Delta \right), i_x \in \mathbb{Z}, i_z \in \mathbb{Z} \right\}. \tag{13}$$

Taking the union of the two discrete half-hyperboloids Λ'_{n,τ_n} and Λ''_{n,τ_n} ensures that the x axis will also eventually have spatial resolution Δ. After the transformation into the coordinate system (x, y, z), we obtain the half-hyperboloid Λ_{n,τ_n} in the search volume G

$$\Lambda_{n,\tau_n} = \{\Omega_n^{-1}\mathbf{R}_n^{-1}(\Lambda'_{n,\tau_n} \cup \Lambda''_{n,\tau_n})\} \cap G. \tag{14}$$

Note that due to the rounding operator, there are regions where two or more hyperboloids corresponding to different TDOAs may be mapped on the same point of the grid. Thus, in contrast to the URG case in which, due to Eq. (4), there are always exactly N TDOA values associated with each point on the grid (one for each microphone pair), the GSG procedure may be associated with less than N, N or more than N TDOAs to a point on the grid. This property is illustrated in Fig. 1, for a section of the search space corresponding to a simulated acoustic environment.

We build the grid map with resolution Δ for all N microphone pairs and for each pair considering all $2T_n + 1$ TDOA values. The values of the discrete hyperboloid and the TDOA information are stored in four look-up tables. We have a table γ_r for the position, a table γ_n for the pair index, and a table γ_τ for the TDOA. For each discrete hyperboloid point $\mathbf{r}_g \in \Lambda_{n,\tau_n}$, the values $\mathbf{r}_g$, n, and τ_n are stored in γ_r, γ_n, and γ_τ respectively. The tables are used in the SRP calculation for estimating the acoustic energy and computing the accumulation of GCC-PHAT functions by all considered sensor pairs. The last look-up table, which we name $\delta(\mathbf{r}_g)$, contains the actual number of surfaces intersecting at position $\mathbf{r}_g$. Specifically, the table $\delta(\mathbf{r}_g)$ is the sensitivity map that gives information on how all sampled GCC-PHAT values are projected into space. In this way, we can obtain a power response sensitivity measure of the considered grid. It will be shown in the experimental section that an improvement of the localization accuracy is obtained in the high sensitivity regions, where the accumulation of GCC-PHAT information is higher. Hence,

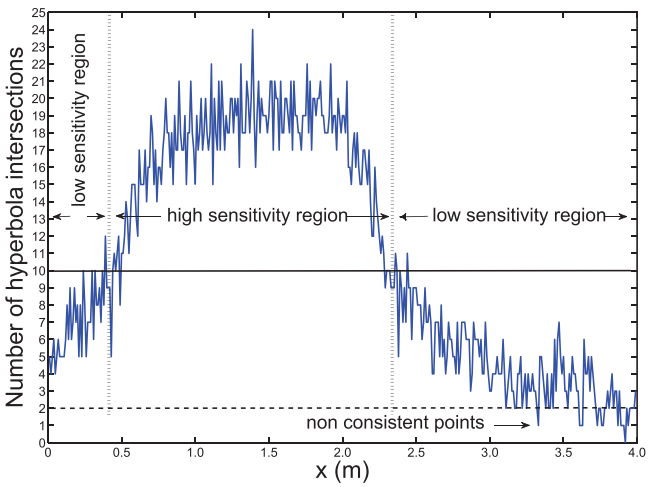

FIG. 1. (Color online) The sensitivity response measure along x axes for a Δ of 0.01 m and $y = 1$ m. The horizontal solid line represents the number of hyperbola intersections assumed by the URG (10 if the number of sensors is 5 as in this case), and the horizontal dashed line represent the minimum number of intersections for acoustical consistency (2 for 2D localization as in this case).

$\forall \mathbf{r}_g \in \Lambda_{n,\tau_n}$, we update the sensitivity map in the following way:

$$\delta(\mathbf{r}_g) = \delta(\mathbf{r}_g) + 1. \tag{15}$$

To be consistent with the definition of a candidate source position as the intersection of hyperboloids, the following constraint is applied after the complete analysis of $\delta(\mathbf{r}_g)$:

$$\forall \mathbf{r}_g \in G, \delta(\mathbf{r}_g) < \mu \Rightarrow \delta(\mathbf{r}_g) = 0, \tag{16}$$

where $\mu = 3$ and $\mu = 2$ in the case of 3D and 2D localization, respectively. The constraint has the goal to discard those space grid points that are not usable for the localization. These grid points are eliminated from the look-up tables γ_r, γ_n, and γ_τ so that all information on the coherent grid representing the relationship with TDOAs of all pair sensor can be used for the localization. Finally, the coherent grid Γ_r related to the array is calculated as

$$\Gamma_r = \{\mathbf{r}_g : \delta(\mathbf{r}_g) \neq 0\}. \tag{17}$$

Figure 2 shows a discrete hyperbola related to a TDOA $t_n = -90$ samples of a specific microphone pair n. The spatial resolution is $\Delta = 0.1$ m, and the area of analysis is $G_x = 4$ m and $G_y = 3$ m. The small circles are the identified grid positions. The grid position $\mathbf{r}_g$, n, and τ_n are stored in the entries γ_r, γ_n, γ_τ. Next, the sensitivity map $\delta(\mathbf{r}_g)$ is updated for each grid point $\mathbf{r}_g$.

The procedure to build the coherently sampled grid and the sensitivity map in a geometric way is given by the following steps.

(1) Initialization of $\delta(\mathbf{r}_g) = 0$ for all $\mathbf{r}_g \in G$;

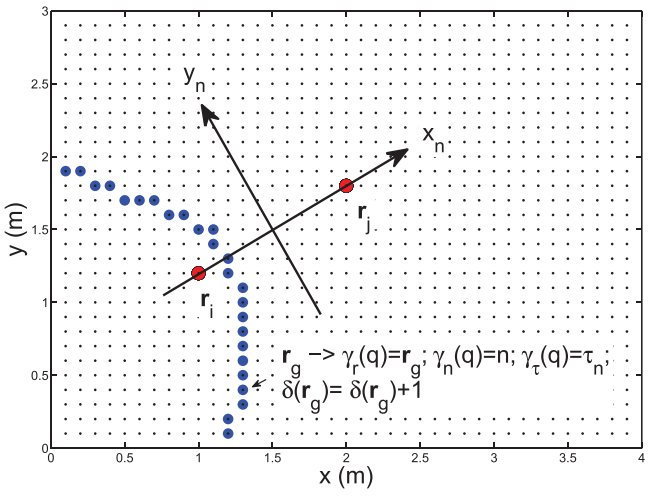

FIG. 2. (Color online) A discrete hyperbola related to TDOA $\tau_n = -90$ samples using the GSG algorithm for a microphone pair $\mathbf{r}_i = [1, 1.2]^T$ m and $\mathbf{r}_j = [2, 1.8]^T$ m. For each grid sample position $\mathbf{r}_g$ of the hyperbola, the values $\mathbf{r}_g$, n, and τ_n are stored in look-up tables γ_r, γ_n, and γ_τ, respectively, and the number of hyperbolas passing through position $\mathbf{r}_g$ are stored in $\delta(\mathbf{r}_g)$. Space resolution Δ is 0.1 m and $f_s = 44.1$ kHz.

(2) For each sensor pair $n = 1, 2, \ldots, N$ and for all TDOA values τ_n in the range $[-T_n, T_n]$, calculate the discrete hyperboloid Λ_{n,τ_n}, and $\forall \mathbf{r}_g \in \Lambda_{n,\tau_n}$ update the value of the sensitivity map $\delta(\mathbf{r}_g) = \delta(\mathbf{r}_g) + 1$ and write the values in the look-up tables γ_r, γ_n, and γ_τ;

(3) After the geometric discrete analysis of hyperboloids has terminated, apply the constraint on $\delta(\mathbf{r}_g)$, update the look-up tables γ_r, γ_n, and γ_τ by removing non-coherent grid points, and calculate Γ_r.

The GSC algorithm is summarized below:

Parameters

N: number of microphone pairs

Δ: spatial resolution

G: search volume

Initialization

$\forall \mathbf{r}_g \in G$, $\delta(\mathbf{r}_g) = 0$

Algorithm

for $n = 1, 2, \ldots, N$, and $\tau_n = -T_n, \ldots, T_n$ **do**

$\Lambda'_{n,\tau_n} = \{(x'_n, y'_n, z'_n) \in \mathbb{R}^3 : (x'_n = \left\lceil \frac{f_x^{\tau_n}(i_y\Delta)}{\Delta} \right\rfloor \Delta, y'_n = \pm \left\lceil \frac{\sqrt{(i_x\Delta)^2 - (i_z\Delta)^2}}{\Delta} \right\rfloor \Delta,$
$z'_n = i_z\Delta), i_y \in \mathbb{Z}, i_z \in \mathbb{Z})\}$

$\Lambda''_{n,\tau_n} = \{(x''_n, y''_n, z''_n) \in \mathbb{R}^3 : (x''_n = i_x\Delta,$
$y''_n = \pm \left\lceil \frac{\sqrt{(f_y^{\tau_n}(i_x\Delta))^2 - (i_z\Delta)^2}}{\Delta} \right\rfloor \Delta, z''_n = i_z\Delta), i_x \in \mathbb{Z}, i_z \in \mathbb{Z}\}$

$\Lambda_{n,\tau_n} = \{\Omega_n^{-1}\mathbf{R}_n^{-1}(\Lambda''_{n,\tau_n} \cup \Lambda''_{n,\tau_n})\} \cap G$

$\forall \mathbf{r}_g \in \Lambda_{n,\tau_n}, \delta(\mathbf{r}_g) = \delta(\mathbf{r}_g) + 1$, update look-up tables $\gamma_r, \gamma_n, \gamma_\tau$

end for

Sensitivity Map

$\forall \mathbf{r}_g \in G, \delta(\mathbf{r}_g) < \mu \Rightarrow \delta(\mathbf{r}_g) = 0$

update look-up tables $\gamma_r, \gamma_n, \gamma_\tau$, and remove entries corresponding to $\delta(\gamma_r) = 0$

Coherent Spatial Grid

$\Gamma_r = \{\mathbf{r}_g : \delta(\mathbf{r}_g) \neq 0\}$

IV. STEERED RESPONSE POWER ALGORITHM USING GSG

The SRP beamformer for source localization is based on the computation of a filtered combination of the signals sensed by the array, upon compensation of their relative phase differences by processing each array channel by an opportune time shift. Typically, a broadband SRP beamformer is computed in the frequency-domain by applying a short-time Fourier transform and by calculating the response power on each frequency bin. Subsequently, a fusion of these estimates is computed. The frequency-domain narrowband output signal of a delay and sum beamforming[21] can be expressed as

$$Y(f, \mathbf{r}_g) = \mathbf{a}^H(f, \mathbf{r}_g)\mathbf{x}(f), \tag{18}$$

where f is the frequency index, the superscript H represents the Hermitian transpose, and $\mathbf{a}(f, \mathbf{r}_g)$ is the steering vector corresponding to a given position $\mathbf{r}_g$.

$\mathbf{x}(f) = [X_1(f), X_2(f), \ldots, X_M(f)]^T$, $Y(f, \mathbf{r}_g)$ and $X_m(f)$, $m = 1, 2, \ldots, M$, are the DFT of the signals. A formal way to express the SRP-PHAT using the beamforming notation is given by

TABLE I. Comparison of number of grid points Γ_G for a ULA using URG and GSG algorithm.

		URG (M = 3,4,5,6)	GSG (M = 3)	GSG (M = 4)	GSG (M = 5)	GSG (M = 6)
$f_s = 16\,000$ Hz	$\Delta = 0.01$ m	40 000 (100%)	486 (1.22%)	3930 (9.83%)	10 854 (27.14%)	20 242 (50.61%)
	$\Delta = 0.05$ m	1600 (100%)	264 (16.50%)	1140 (71.25%)	1446 (90.38%)	1509 (94.31%)
	$\Delta = 0.1$ m	400 (100%)	185 (46.25%)	358 (89.50%)	370 (92.50%)	374 (93.50%)
$f_s = 44\,100$ Hz	$\Delta = 0.01$ m	40 000 (100%)	3710 (9.28%)	15 816 (39.54%)	29 708 (74.27%)	36 958 (92.40%)
	$\Delta = 0.05$ m	1600 (100%)	1281 (80.06%)	1527 (95.44%)	1540 (96.25%)	1559 (97.44%)
	$\Delta = 0.1$ m	400 (100%)	372 (93.00%)	378 (94.50%)	380 (95.00%)	380 (95.00%)
$f_s = 96\,000$ Hz	$\Delta = 0.01$ m	40 000 (100%)	12 362 (30.91%)	31 908 (79.77%)	38 358 (95.90%)	39 103 (97.76%)
	$\Delta = 0.05$ m	1600 (100%)	1512 (94.50%)	1535 (95.94%)	1548 (96.75%)	1552 (97.00%)
	$\Delta = 0.1$ m	400 (100%)	374 (93.50%)	380 (95.00%)	380 (95.00%)	380 (95.00%)

$$P(\mathbf{r}_g) = \sum_{f=0}^{L-1} E\{|Y(f,\mathbf{r}_g)|^2\}, \\ = \sum_{f=0}^{L-1} \mathbf{a}^H(f,\mathbf{r}_g)(\Phi(f) \div |\Phi(f)|)\mathbf{a}(f,\mathbf{r}_g), \tag{19}$$

where $P(\mathbf{r}_g)$ is the power spectral density of the beamformer output in position $\mathbf{r}_g$, L is the length of the DFT analysis window, $E\{\cdot\}$ denotes mathematical expectation, $\Phi(f) = E\{\mathbf{x}(f)\mathbf{x}^H(f)\}$ is the cross-spectral density matrix, $\div$ denotes element-wise division, and $|\cdot|$ denotes the element-wise absolute value operation. The PHAT filter discards the magnitude and only keeps the phase of $\Phi(f)$ for computing the normalized steered responses.

The steered response power can be efficiently implemented using the GCF and GCC-PHAT functions.[5] Therefore, we have that the SRP-PHAT with the URG becomes

$$P_{\mathrm{URG}}(\mathbf{r}_g) = \sum_{n=1}^{N} R_n[\tau_n(\mathbf{r}_g)], \tag{20}$$

where the GCC using the PHAT whitening for a generic n pair is given by

$$R_n[\tau] = \frac{1}{L}\sum_{f=0}^{L-1} \frac{X_i(f)X_j^*(f)}{|X_i(f)X_j^*(f)|} e^{j2\pi f\tau/L}, \tag{21}$$

where τ is the time lag and $(\cdot)^*$ denotes the complex conjugate.

The equation of the SRP-PHAT power spectral density computed using the GSG algorithm, although similar to Eq. (20), takes into account all the discrete TDOA values and the acoustically coherent space grid points contained in the look-up tables $\gamma_r, \gamma_n, \gamma_\tau$:

$$P_{\mathrm{GSG}}(\mathbf{r}_g) = \sum_{h\in H_r} R_{\gamma_n(h)}[\gamma_\tau(h)], \tag{22}$$

where

$$H_r = \{q : \gamma_r(q) = \mathbf{r}_g\} \tag{23}$$

is the set of look-up table indices corresponding to the TDOAs of all the N sensor pairs for the position $\mathbf{r}_g \in \Gamma_r$. Note that H_r is a set of TDOAs of dimension $\delta(\mathbf{r}_g)$. After some manipulation of Eq. (22), we can write the SRP-PHAT-GSG as

$$P_{\mathrm{GSG}}(\mathbf{r}_g) = \sum_{n=1}^{N} \sum_{z\in Z_{r,n}} R_n[\gamma_\tau(z)], \tag{24}$$

where

$$Z_{r,n} = \{q : [\gamma_r(q) = \mathbf{r}_g] \wedge [\gamma_n(q) = n]\} \tag{25}$$

is the look-up table indices corresponding to the TDOAs for the position $\mathbf{r}_g \in \Gamma_r$ of the sensor pair n. Note that $Z_{r,n}$ is an empty set if $\{q : [\gamma_r(q) = \mathbf{r}_g] \wedge [\gamma_n(q) = n]\}$ is null. By comparing Eqs. (20) and (24), we can observe that for each

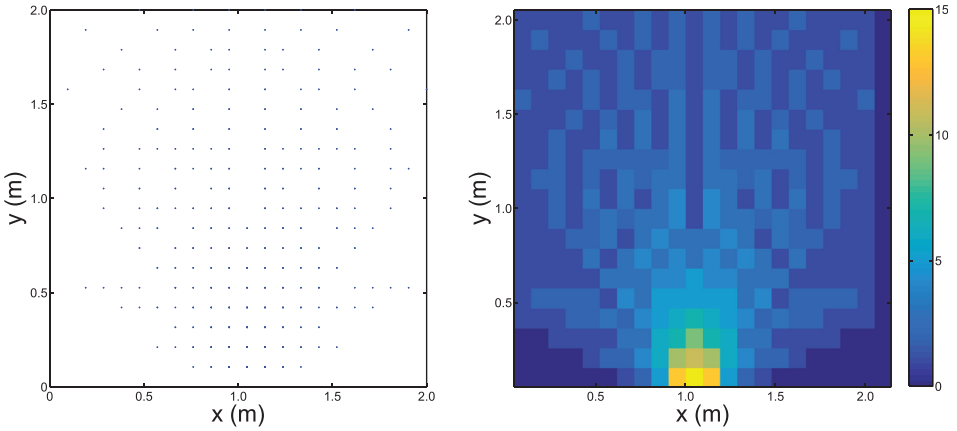

FIG. 3. (Color online) The grid map Γ_r and the sensitivity map $\delta(\mathbf{r}_g)$ for an ULA of three microphones, a space resolution $\Delta = 0.1$ m, and $f_s = 16$ kHz.

position related to the microphone pair n, we can have a larger amount of TDOA information, which is the principal reason of the increased localization performance in the high sensitivity region. Note that the SRP-PHAT expressed by Eq. (24) has a similar form of other accumulation methods.[16–18] However, GSG designs a coherent spatial grid and provides a sensitivity map, which gives information of how the whole GCC-PHAT information is distributed in the search space, resulting in different regions characterized by different localization accuracies.

For an analysis frame at time k composed of L samples, the GSG based SRP-PHAT is computed in three steps. First, the map is initialized by imposing the steered response power $P_{\mathrm{GSG}}(\mathbf{r}_g) = 0$ with $\mathbf{r}_g \in \Gamma_r$. Then, the values from the estimated GCC-PHAT functions are accumulated in the grid map. Finally, the source position is estimated by picking the maximum value of the acoustic map

$$\hat{\mathbf{r}}_s(k) = \underset{\mathbf{r}_g}{\operatorname{argmax}}\left[P_{\mathrm{GSG}}(\mathbf{r}_g)\right], \mathbf{r}_g \in \Gamma_r. \tag{26}$$

The SRP-PHAT-GSG is summarized below:

Initialization

$\forall \mathbf{r}_g \in \Gamma_r,\ P_{\mathrm{GSG}}(\mathbf{r}_g) = 0$

Algorithm

$$P_{\mathrm{GSG}}(\mathbf{r}_g) = \sum_{h \in H_r} R_{\gamma_n(h)}[\gamma_\tau(h)], H_r = \{q : \gamma_r(q) = \mathbf{r}_g\}$$

$$\hat{\mathbf{r}}_s(k) = \underset{\mathbf{r}_g}{\operatorname{argmax}}[P_{\mathrm{GSG}}(\mathbf{r}_g)], \mathbf{r}_g \in \Gamma_r$$

The computational cost for the GSG algorithm is equivalent to that of the URG procedure for computing the power map, since for both algorithms the relationship between TDOAs and positions in space is pre-calculated offline using the look-up tables, and online summation is negligible. The major computational demand of SRP-PHAT-GSG comes from the number of grid points Γ_G, and hence the complexity is $\mathcal{O}(\Gamma_G)$. On the other hand, the major computational demand of indirect methods comes from LS operations. For example, the constrained LS (CLS) criterion in Ref. 3 requires two matrix inversion operations, and the complexity is $\mathcal{O}(2N_p^3)$, where N_p is the number of considered microphone pairs. For the proposed GSG, a consistent reduction may occur for the search procedure computational cost, which depends on the number of sample grid positions. If the search procedure is restricted to the coherent grid, the computational cost is inferior to the URG method due to the discarded points. Moreover, the computational cost may also be reduced by using a coarser grid or by only searching in the high sensitivity regions, in which the localization accuracy is maximized.

V. EXPERIMENTAL RESULTS

A. Spatial grid and power response sensitivity analysis

In this section, we present experimental results concerning the construction of the spatial grid and the analysis of the power response sensitivity using the GSG algorithm for an uniform linear array (ULA). Spatial grids were designed using different small-array sizes, sampling rate values, and spatial resolutions. A search region of 2 m × 2 m was considered. Table I shows the resulting number of grid points Γ_G when using the URG and the GSG methods, for an ULA with an inter-microphone distance of 0.15 m. The coverage percentage values reported show how the acoustically coherent grid is in some cases much smaller if compared to the uniform regular grid (especially when using a small array size combined with a high spatial resolution). As already noted, using the coherent spatial grid obtained by the GSG algorithm in those cases has the advantage of providing a position search domain which is consistent with the hyperboloid intersections, whereas the URG grid would also contain non-consistent regions which would provide misleading information, since the corresponding energy on the search map is usually comparable to that of consistent regions.

Figures 3–8 depict the grid map Γ_r and the sensitivity map $\delta(\mathbf{r}_g)$ calculated with the GSG algorithm for different system configurations. The center of the array is positioned at location (1,0) m. Note that the $\delta(\mathbf{r}_g)$ tables in the figures are reported before applying the constraint in Eq. (16). The bar on the right of the figures shows the number of the intersections of hyperbolas.

By observing the sensitivity maps, we can see how the GCC-PHAT functions are projected onto the search region,

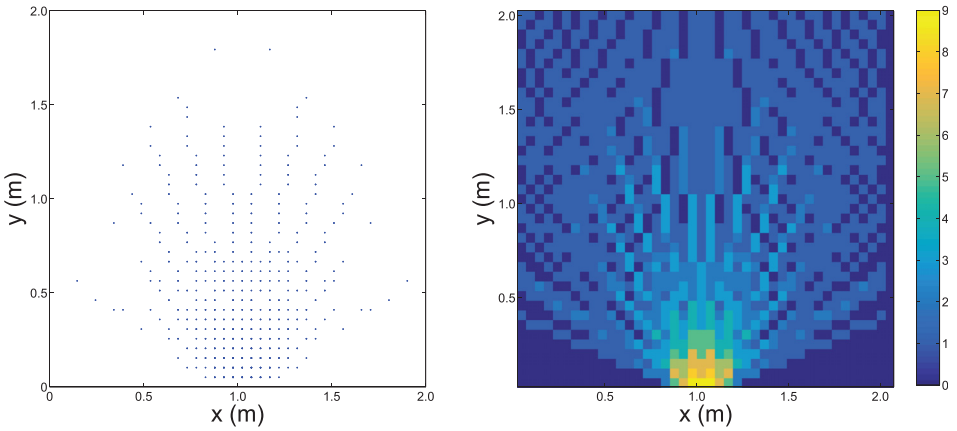

FIG. 4. (Color online) The grid map Γ_r and the sensitivity map $\delta(\mathbf{r}_g)$ for an ULA of three microphones, a space resolution $\Delta = 0.05$ m, and $f_s = 16$ kHz.

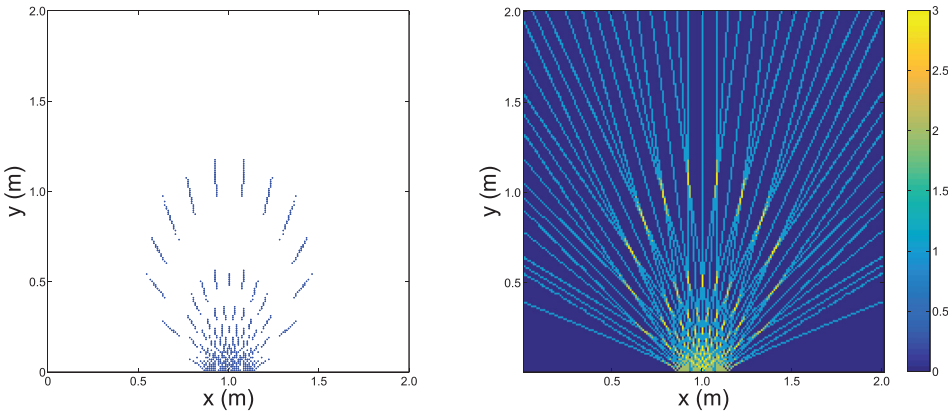

FIG. 5. (Color online) The grid map Γ_r and the sensitivity map $\delta(\mathbf{r}_g)$ for an ULA of three microphones, a space resolution $\Delta = 0.01$ m, and $f_s = 16$ kHz.

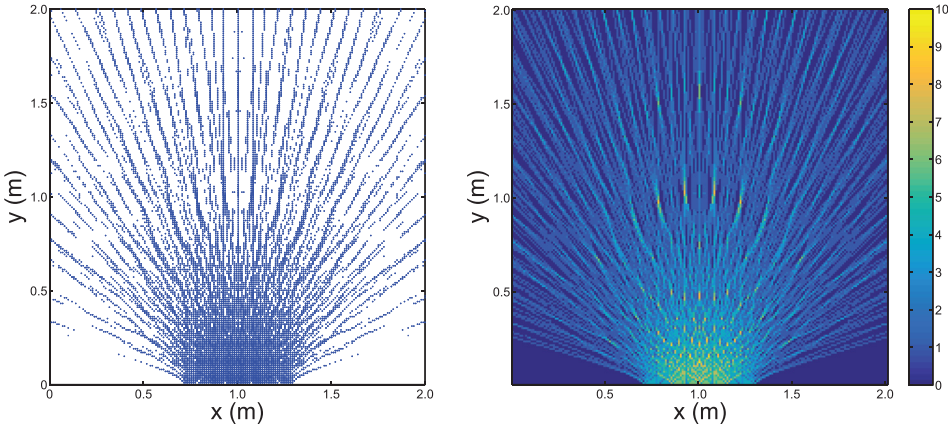

FIG. 6. (Color online) The grid map Γ_r and the sensitivity map $\delta(\mathbf{r}_g)$ for an ULA of five microphones, a space resolution $\Delta = 0.01$ m, and $f_s = 16$ kHz.

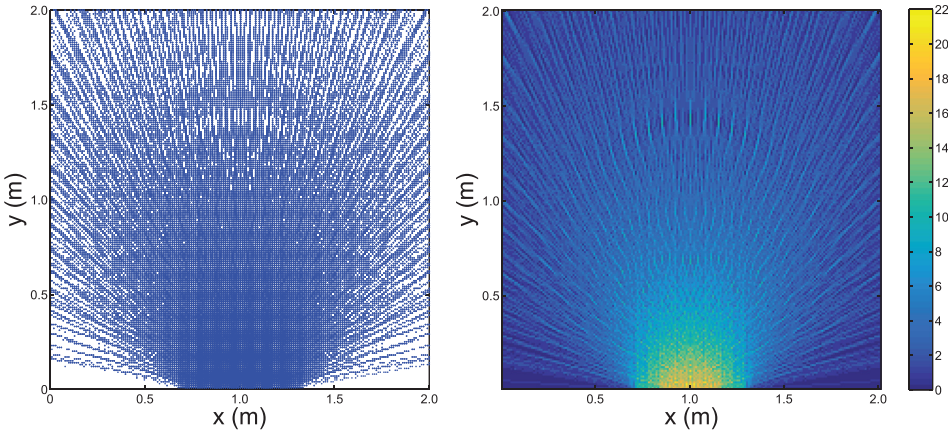

FIG. 7. (Color online) The grid map Γ_r and the sensitivity map $\delta(\mathbf{r}_g)$ for an ULA of five microphones, a space resolution $\Delta = 0.01$ m, and $f_s = 44.1$ kHz.

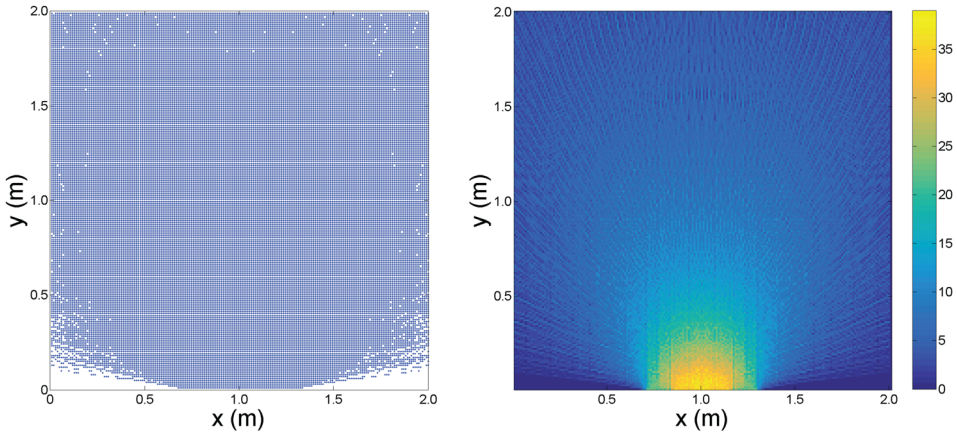

FIG. 8. (Color online) The grid map Γ_r and the sensitivity map $\delta(\mathbf{r}_g)$ for an ULA of five microphones, a space resolution $\Delta = 0.01$ m, and $f_s = 96$ kHz.

and how their values are accumulated. We note that the high value regions are characterized by a high power response sensitivity since they accommodate a high number of hyperbola intersections. We can see in Fig. 8 that the high sensitivity region accommodates a number of intersections contained in the range [25,35] whereas the URG only accounts for $M(M-1)/2 = 10$ intersections at each point on the grid. Figure 9 depicts the power response sensitivity analysis corresponding to different values of the array aperture, for an ULA of five microphones, a space resolution $\Delta = 0.01$ m, and $f_s = 96$ kHz. We observe how the high sensitivity region expands when the distance between microphone increases, due to the higher resolution of the GCC-PHAT functions that provide a larger number of hyperbolas for each sensor pair.

The coherent spatial grid and the sensitivity map can be optimally constructed for a specific search region by properly configuring the geometry of the array, the number of microphones, and the sampling frequency. An alternative way to increase the TDOA resolution, and accordingly the number of hyperboloid of a sensor pair, is by interpolation of GCC-PHAT functions. If $1/\alpha$ is an upsampling step, the possible TDOA values for the sensor pair n will become $2\alpha T_n + 1$. When GCC-PHAT interpolation is considered in the GSG, we also have to calculate discrete hyperboloids for non-integer TDOA values according to the parameter α. An example of grid maps in the GSG is shown in Fig. 10, in which we can observe the spatial grid corresponding to different values of α, for an ULA of four microphones, a space resolution $\Delta = 0.01$ m, and $f_s = 8$ kHz. Note that the effectiveness of interpolation of GCC-PHAT functions for incrementing the spatial resolution is related to the signal-to-noise ratio (SNR) of the signal, and upsampling may lead to poor accuracy for low SNR.[22]

In the following, we will see the importance of the power response sensitivity analysis and how it is deeply related to the performance of sound source localization.

B. Localization performance for simulated data

In this section, the localization performance of the proposed GSG algorithm is assessed on a set of acoustic data simulated numerically. We also show that the sensitivity map obtained with the GSG algorithm is a useful tool to classify the areas in terms of high or poor localization performance. Besides that, we compare the performance of SRP-PHAT using URG,[5] URG-H,[16] URG-M,[17] URG-I,[18] and GSG algorithm for different spatial resolution conditions: low $\Delta = 0.5$ m, medium $\Delta = 0.05$ m, and high $\Delta = 0.01$ m. We also consider the indirect method based on GCC-PHAT (Ref. 10) and CLS.[3]

In the experiments with simulated acoustic data, a randomly distributed microphone network of five sensors was used. The image-source method (ISM) was used to simulate reverberant audio data in room acoustics.[23,24] A localization task in two-dimensions, in a room of $4\,\text{m} \times 3\,\text{m} \times 3\,\text{m}$, was considered. Therefore, both microphones and the source were positioned at a distance from the floor of 1.7 m. The room setup is shown in Fig. 11.

The δ table calculated with the GSG algorithm for Δ's of 0.01 m, 0.05 m, and 0.5 m are depicted in Figs. 12–14,

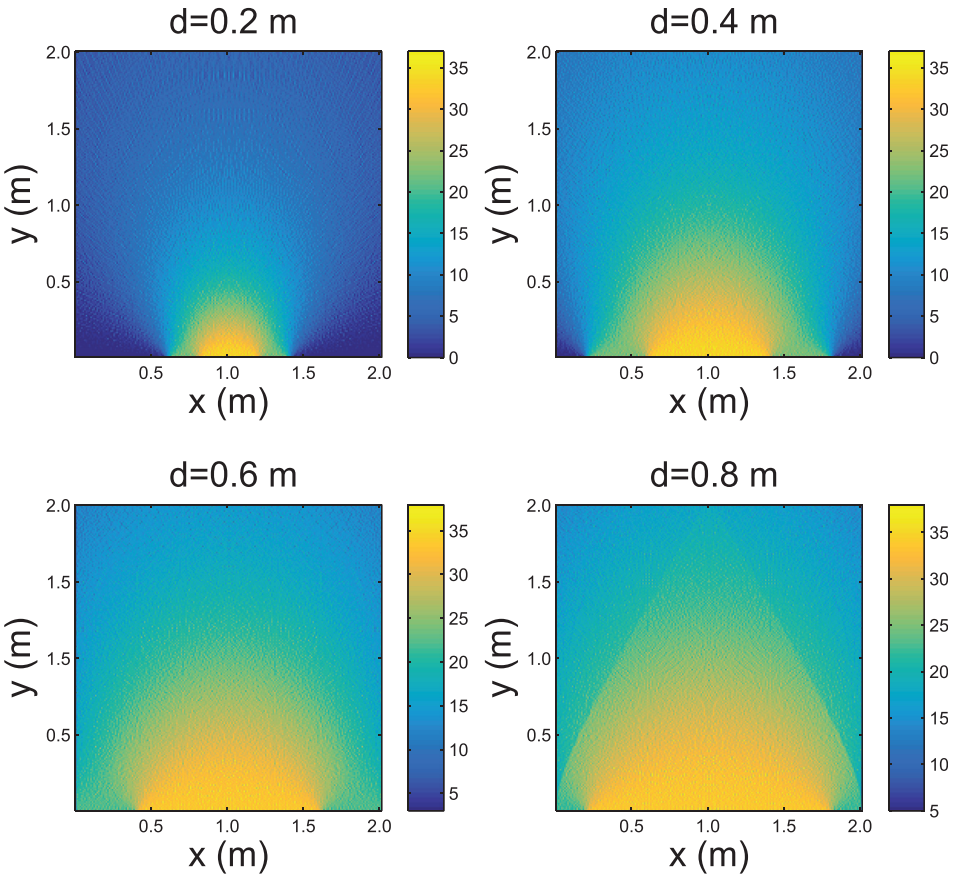

FIG. 9. (Color online) The sensitivity map $\delta(\mathbf{r}_g)$ corresponding to four values of the inter-microphone distance d for an ULA of five microphones, a space resolution $\Delta = 0.01$ m, and $f_s = 96$ kHz.

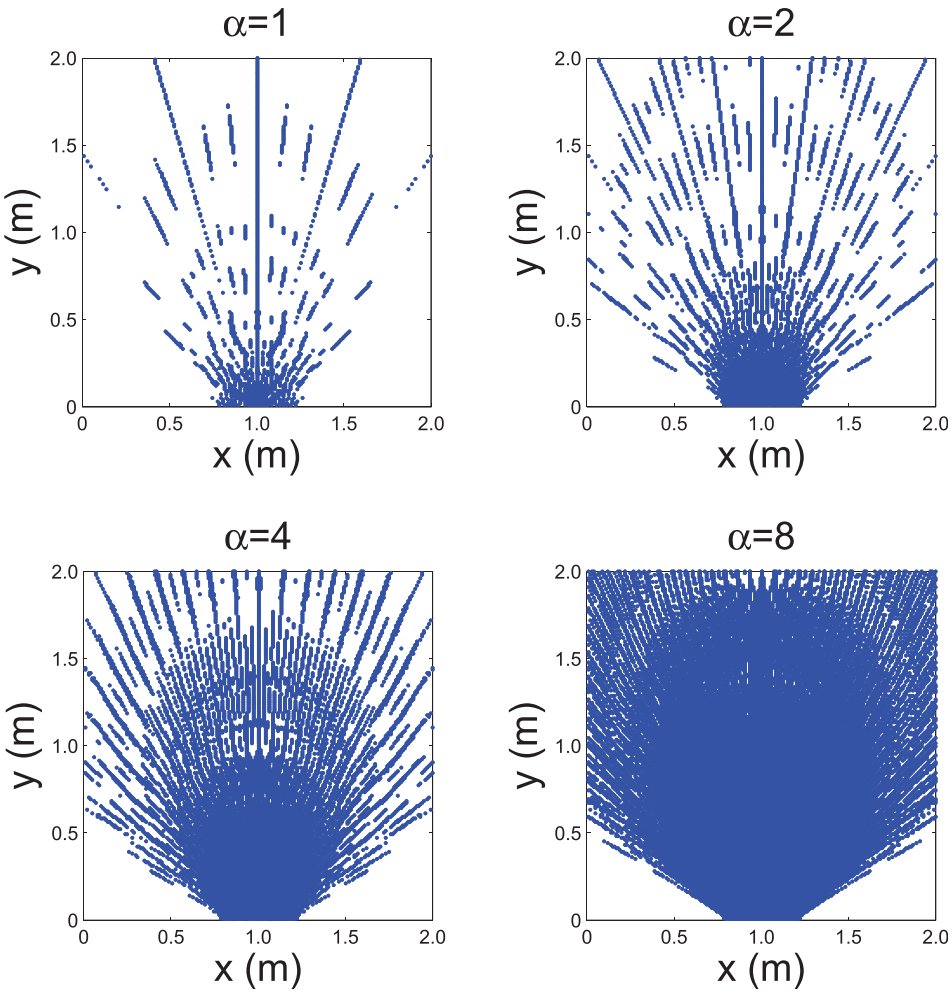

FIG. 10. (Color online) The grid map Γ_r with ($\alpha = 2,4,8$) and without ($\alpha = 1$) TDOA upsampling, for an ULA of four microphones, a space resolution $\Delta = 0.01$ m, and $f_s = 8$ kHz.

respectively. We also report the discriminability measure map proposed in Ref. 20. As we can observe in Figs. 15–17 the discriminability measure map is accurate for $\Delta = 0.01$ m but it does not provide useful information for $\Delta = 0.05$ m and $\Delta = 0.5$ m, because of the TDOA information loss discussed so far. Figure 1 shows the sensitivity response measure in terms of hyperbola intersections along x axes for a Δ of 0.01 m and $y = 1$ m. The horizontal solid line represents the number of hyperbola intersections assumed by the URG. We note a greater number of intersections in the high sensitivity region with a range $x = [0.4, 2.3]$ m.

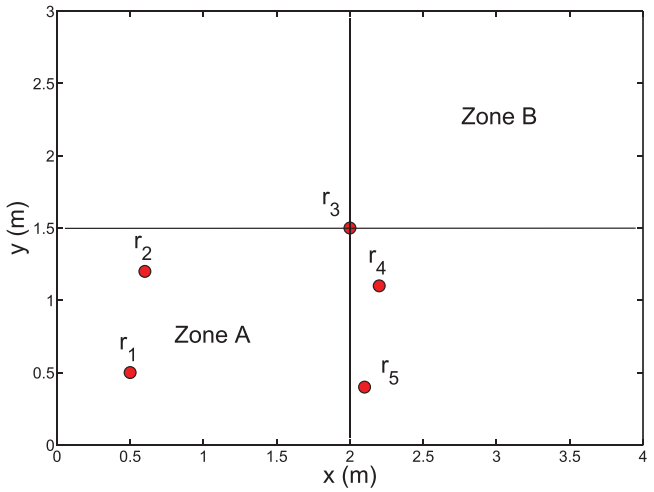

FIG. 11. (Color online) The simulated room setup with the positions of the five microphones and the two zones A and B for evaluating the performance of SRP-PHAT with URG, URG-I, URG-M, URG-H, and GSG algorithms. Two zones A and B were considered with high and low TDOA information taking into account the sensitivity map depicted in Figs. 12, 13, and 14.

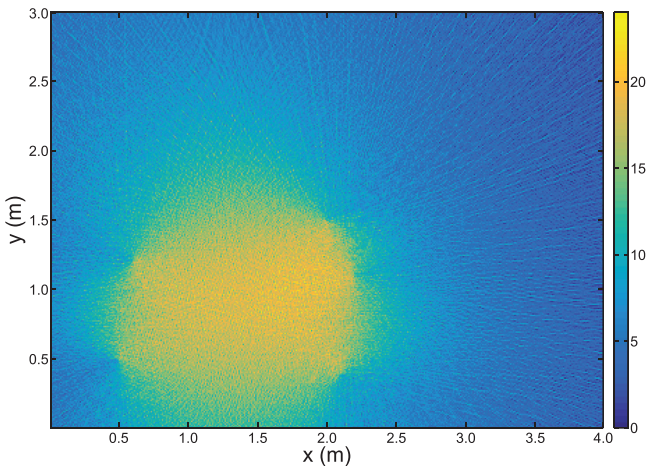

FIG. 12. (Color online) The sensitivity map $\delta(\mathbf{r}_g)$ provided by the GSG of the array in Fig. 10 with $\Delta = 0.01$ m.

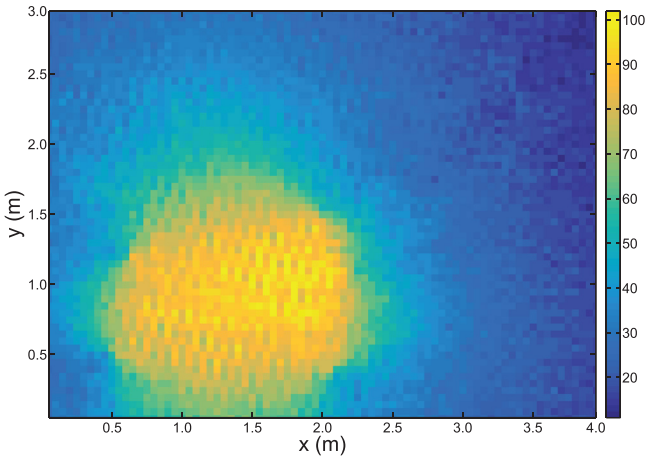

FIG. 13. (Color online) The sensitivity map $\delta(\mathbf{r}_g)$ provided by GSG of the array in Fig. 10 with $\Delta = 0.05$ m.

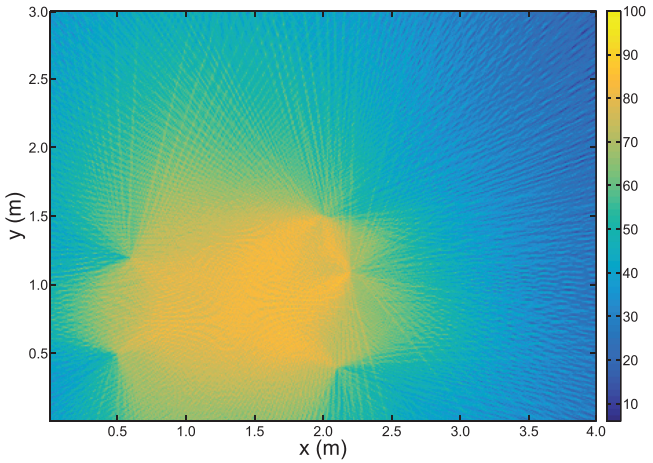

FIG. 15. (Color online) The discriminability measure map (Ref. 20) of the array in Fig. 10 with $\Delta = 0.01$ m.

The reverberant condition was set to 0.3 and 0.9 s reverberation time (RT_{60}). A 25 s duration adult male speech was used as a source signal. The tests were conducted by setting a SNR of 10 dB, which was obtained by adding mutually independent white Gaussian noise to each channel. The sampling frequency was 44.1 kHz, the block size L was 4096 samples.

Two zones A and B were considered with high and low TDOA information, taking into account the sensitivity map depicted in Figs. 12–14. The performance of localization has been evaluated with several Monte Carlo simulations, using 100 run-trials for each condition test. The source was randomly positioned at each trail, at a minimum distance of 0.1 m from the walls and microphones. Performance is reported in terms of the percentage of accuracy rate (AR) estimated for those square errors that are less than a root mean square error (RMSE) of 0.2 m, and by the RMSE for all the estimates.

The localization performance is given in Table II. First, we can observe that SRP-PHAT-GSG outperforms SRP-PHAT-URG in all test conditions for zone A. Besides that, we note a rapid degradation of SRP-PHAT-URG performance when the spatial resolution decreases, while SRP-PHAT-GSG is more robust due to the improved TDOA information exploitation. Then, we have that the number of grid points for GSG is the same of URG when $\Delta = 0.1$ m ($\Gamma_G = 48$) and $\Delta = 0.05$ m ($\Gamma_G = 1200$). However, in the case of $\Delta = 0.01$ m the GSG grid points are about 3% less than the URG grid points, slightly reducing the computational cost for the maximum value search. The average performance of the URG-M and URG-H is comparable to that of the GSG. Specifically, GSG has a better AR and RMSE in coarser grids ($\Delta = 0.1$ and 0.05 m), due to the use of all TDOA information that ensures a larger number of hyperbola intersections in the high sensitivity region. URG-M and URG-H provide instead better performance when $\Delta = 0.01$ m. In this case, the use of a fine grid reduces the accumulation of GSG. However, URG-M and URG-H provide no clues to select the region with best localization accuracy, while GSG includes the sensitivity analysis, which gives important clues on how all of the TDOA information is distributed. In fact, in the low accuracy zone B, all algorithms perform the localization with a higher error if compared to zone A. When reverberation time increases, the noisier condition degrades the GCC-PHAT performance and the poor

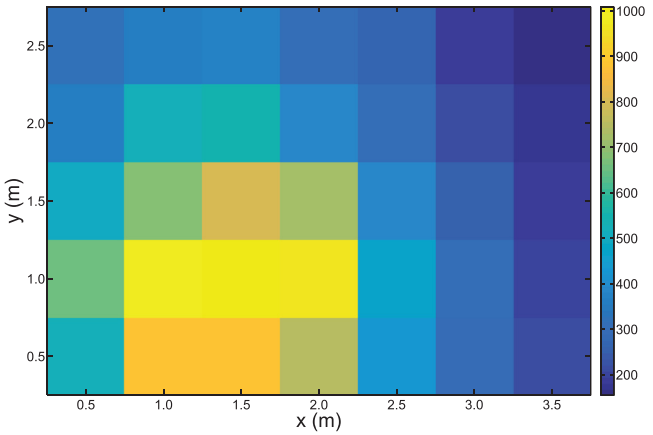

FIG. 14. (Color online) The sensitivity map $\delta(\mathbf{r}_g)$ provided by GSG of the array in Fig. 10 with $\Delta = 0.5$ m.

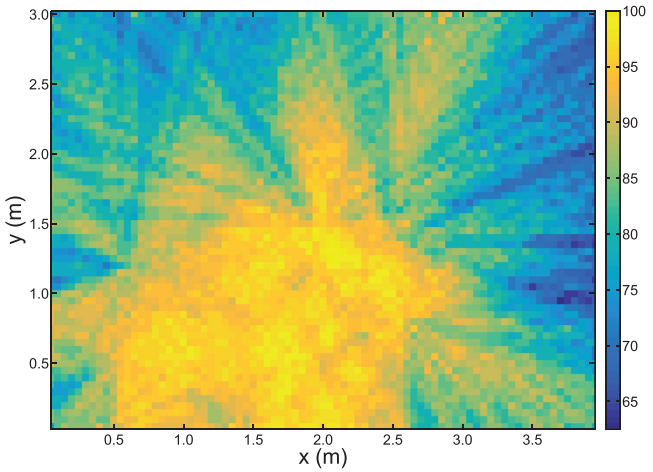

FIG. 16. (Color online) The discriminability measure map (Ref. 20) of the array in Fig. 10 with $\Delta = 0.05$ m.

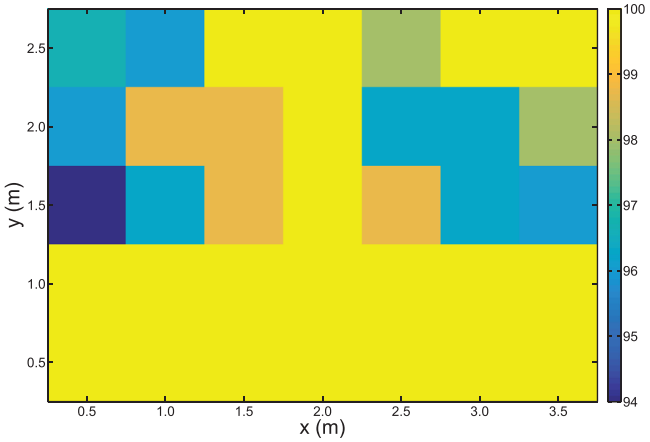

FIG. 17. (Color online) The discriminability measure map (Ref. 20) of the array in Fig. 10 with $\Delta = 0.5$ m.

TDOA information in that region makes the localization very difficult. In particular, GSG, URG-M, and URG-H are affected by a consistent performance degradation due to the fact that in zone B a low energy peak related to the acoustic source is subject to be masked by high energy noise peaks with high probability. This observation suggests that a zone selection procedure that gives information on which is the most promising searching area may help in increasing the localization performance of GSG, URG-M, and URG-H in low level sensitivity zones. The URG-I provides worse localization performance for zone A if compared to that of GSG, URG-M, and URG-H, due to the averaging of the GCC-PHAT for each volume of the search grid. The localization performance of GCC-PHAT CLS is given in Table III. We can observe the worse performance in comparison to SRP-based methods, and the different performance in low and high sensitivity regions.

TABLE III. RMSE (m) and AR (%) (RMSE < 0.2 m) of 2D localization performance for GCC-PHAT with CLS in a simulated reverberant room using a speech signal and a SNR of 10 dB.

			GCC-PHAT CLS
$RT_{60} = 0.3$ s	Zone A	RMSE (m)	2.584
		AR (%)	28.63
	Zone B	RMSE (m)	4.692
		AR (%)	0.07
$RT_{60} = 0.9$ s	Zone A	RMSE (m)	3.381
		AR (%)	4.81
	Zone B	RMSE (m)	4.984
		AR (%)	0.04

Next, we also provide a validation of the GSG in a 3D environment with two planar array geometries that can be used to locate the source in the half-space due to the front-rear ambiguity. A randomly distributed microphone depicted in Fig. 11 and a T-shaped array of six microphones were located in a room of $4\,\text{m} \times 3\,\text{m} \times 3\,\text{m}$ with a RT_{60} of 0.3 s. The randomly distributed microphone was positioned at a distance from the floor of 0.5 m and the volume above the array is considered as the localization area. The T-shaped array was obtained by disposing four microphones horizontally and two microphones in the vertical symmetry plane of the array. The inter-microphone distance was 0.3 m. The high and low sensitivity regions are identified with the threshold value η defined as

TABLE II. RMSE (m) and AR (%) (RMSE < 0.2 m) of 2D localization performance for SRP-PHAT with GSG, URG, URG-I, URG-M, URG-H in a simulated reverberant room using a speech signal and a SNR of 10 dB.

				GSG	URG	URG-I	URG-M	URG-H
$RT_{60} = 0.3$ s	$\Delta = 0.5$ m	Zone A	RMSE (m)	0.600	1.679	1.536	0.668	0.637
			AR (%)	38.76	6.32	12.97	35.55	35.30
		Zone B	RMSE (m)	1.898	1.622	1.476	1.834	1.849
			AR (%)	1.14	3.92	6.19	2.39	1.66
	$\Delta = 0.05$ m	Zone A	RMSE (m)	0.292	1.224	1.564	0.310	0.315
			AR (%)	87.79	48.00	58.67	87.25	86.57
		Zone B	RMSE (m)	2.027	1.496	1.103	1.960	1.969
			AR (%)	6.91	30.29	38.01	13.29	12.75
	$\Delta = 0.01$ m	Zone A	RMSE (m)	0.257	0.665	1.262	0.243	0.229
			AR (%)	90.75	77.80	71.53	91.01	91.68
		Zone B	RMSE (m)	2.112	1.719	1.175	2.028	1.994
			AR (%)	3.56	28.77	35.21	10.12	16.84
$RT_{60} = 0.9$ s	$\Delta = 0.5$ m	Zone A	RMSE (m)	0.795	1.750	1.778	0.867	0.855
			AR (%)	21.83	3.27	4.12	19.87	18.80
		Zone B	RMSE (m)	2.063	1.771	1.775	2.045	2.057
			AR (%)	0.27	2.06	2.70	0.53	0.41
	$\Delta = 0.05$ m	Zone A	RMSE (m)	0.540	1.627	2.230	0.553	0.558
			AR (%)	57.96	16.35	17.42	57.88	57.91
		Zone B	RMSE (m)	2.177	1.917	1.569	2.168	2.170
			AR (%)	1.06	7.95	11.21	2.49	2.34
	$\Delta = 0.01$ m	Zone A	RMSE (m)	0.534	1.139	2.056	0.547	0.531
			AR (%)	61.93	40.86	31.06	62.90	65.32
		Zone B	RMSE (m)	2.138	2.078	1.592	2.122	2.130
			AR (%)	0.52	7.34	10.03	2.65	3.13

TABLE IV. RMSE (m) and AR (%) (RMSE < 0.2 m) of 3D localization performance for GSG in a simulated reverberant room using a speech signal, a RT_{60} of 0.3 s, and a SNR of 10 dB.

			GSG
Randomly distributed microphone array	High sensitivity region	RMSE (m)	0.514
		AR (%)	32.65
	Low sensitivity region	RMSE (m)	2.109
		AR (%)	1.13
T-shaped array	High sensitivity region	RMSE (m)	0.276
		AR (%)	45.31
	Low sensitivity region	RMSE (m)	2.489
		AR (%)	0.14

$$\eta = \frac{\max\left(\delta(\mathbf{r}_g)\right)}{2\left(\max\left(\delta(\mathbf{r}_g)\right) - \min\left(\delta(\mathbf{r}_g)\right)\right)}.$$

The position $\mathbf{r}_g$ belongs to the high sensitivity region if $\delta(\mathbf{r}_g) \geq \eta$ and to the low sensitivity region if $\delta(\mathbf{r}_g) < \eta$. The performance of localization has been evaluated with several Monte Carlo simulations, using 100 run-trials for each zone. The source was randomly positioned at each trail. The tests were conducted with a speech signal, a grid resolution of $\Delta = 0.05$ m, and a SNR of 10 dB. Table IV shows the results for the two arrays and the two zones providing different performance in the low and the high sensitivity regions.

C. Localization performance for real data

We report extensive tests computed in a real-world setup. An acoustic sensor network of 24 microphones has been installed in a conference room equipped with various multimedia facilities. The net of microphones is composed of three arrays, each one composed by eight microphones arranged in a ULA with a distance between sensors of 0.16 m. The arrays are positioned with a distance from the floor of 1.7 m. The room setup is shown in Fig. 18, which also reports the source position (black circles) that has been used during recordings. The room dimensions in the x, y, z coordinates was 16 m × 7 m × 3 m, and its measured reverberation time was approximately 0.9 s of RT_{60}. The high reverberation time is due to the presence of glass window panes on the two sidewalls of the room. We have considered a position search area of dimensions 9.2 m × 3.88 m, and the δ table was calculated with the GSG algorithm for an imposed spatial resolution Δ of 0.05 m. The resulting sensitivity map $\delta(\mathbf{r}_g)$ is depicted in Fig. 19. The grid points calculated with the GSG algorithm cover all the localization area, i.e., they are equal to URG in this specific case. All microphone pairs of each array has been used so that $N = 84$. We have defined two zones (see Fig. 18) for evaluating the localization performance taking into account the sensitivity map depicted in Fig. 19: a high sensitivity region (zone C) and a low sensitivity region (zone D).

A speech database was recorded in the conference room, which consists of short sentences uttered by two male speakers and one female speaker, standing up at different positions in the room showed in Fig. 18 with black circles. The recordings were organized in ten sessions, in which one speaker for each session changed four to eight locations, each time repeating his new position in the room. The total database consists of about 30 min. of audio. The 24-channel audio was acquired at 48 kHz. The SRP-PHAT was computed with a block size L of 4096 samples and an overlap step of $L/4$. The parameters are

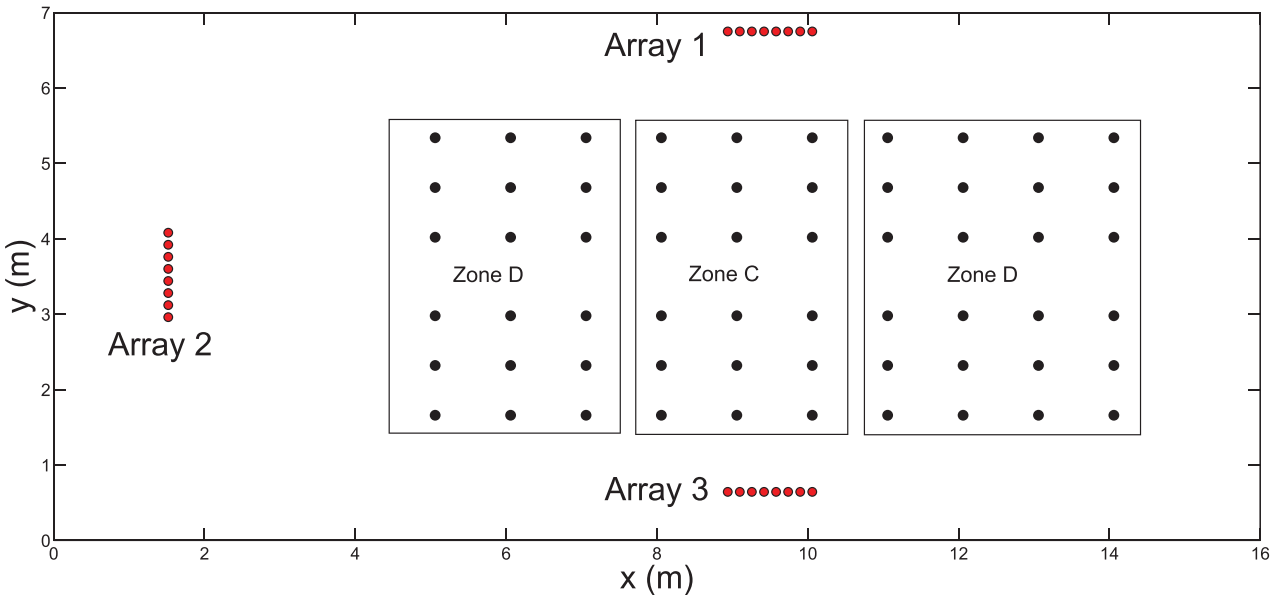

FIG. 18. (Color online) The real-world room setup with the positions of the microphones and the speakers using three linear arrays. Two zones C and D were considered with high and low TDOA information taking into account the sensitivity map depicted in Fig. 19.

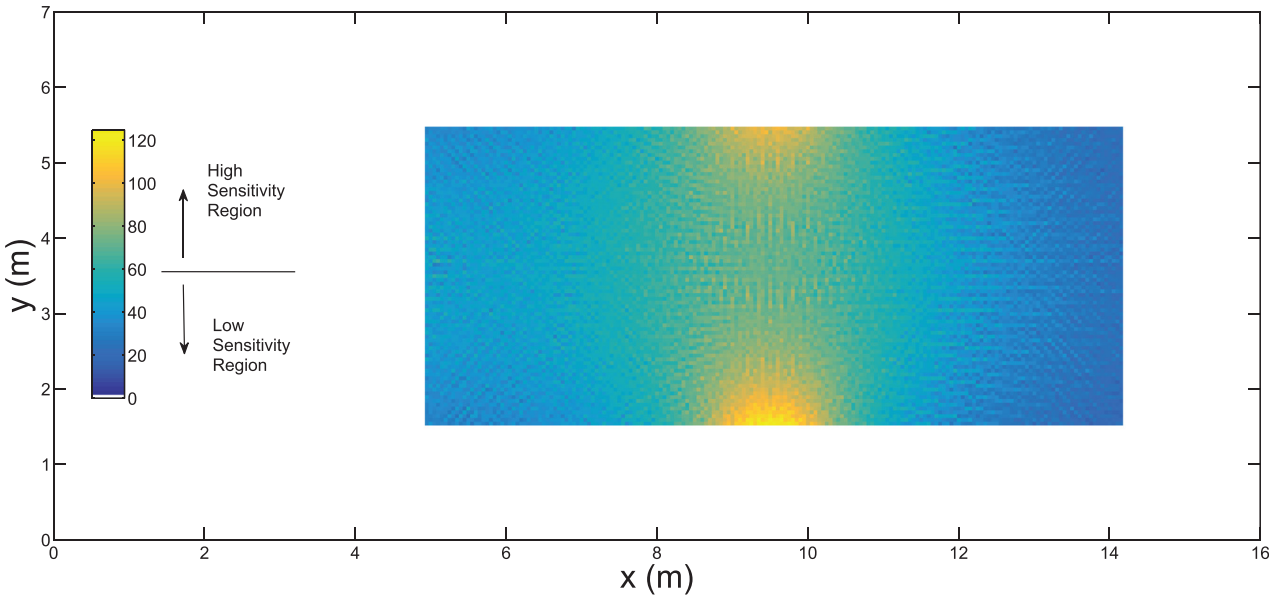

FIG. 19. (Color online) The sensitivity map $\delta(\mathbf{r}_g)$ of the array in Fig. 18 with $\Delta = 0.05$ m and $f_s = 48$ kHz.

TABLE V. RMSE (m) and AR (%) (RMSE < 0.2 m) of localization performance for SRP-PHAT with GSG, URG, URG-I, URG-M, and URG-H in a real room with a RT_{60} of 0.9 s using three linear arrays.

		GSG	URG	URG-I	URG-M	URG-H
Zone C	RMSE (m)	1.267	1.737	1.986	1.134	1.161
	AR (%)	32.42	22.34	22.39	27.53	26.41
Zone D	RMSE (m)	3.428	2.799	3.011	2.789	2.699
	AR (%)	7.65	9.82	10.60	10.06	11.40

evaluated in terms of AR percentage estimates for RMSE < 0.2 m, and overall RMSE.

Table V shows the obtained results for the two zones. As we can see, the localization performance of all algorithms is more robust in terms of RMSE and AR in the high sensitivity region (zone C) and we can observe the decrease of performance of all algorithms when the source was positioned in the low sensitivity region (zone D). Note that the distinction between high-sensitivity and low-sensitivity areas in the search space is less marked than it was in the simulated experiments. Actually, the most of zone C turns out to be characterized by a

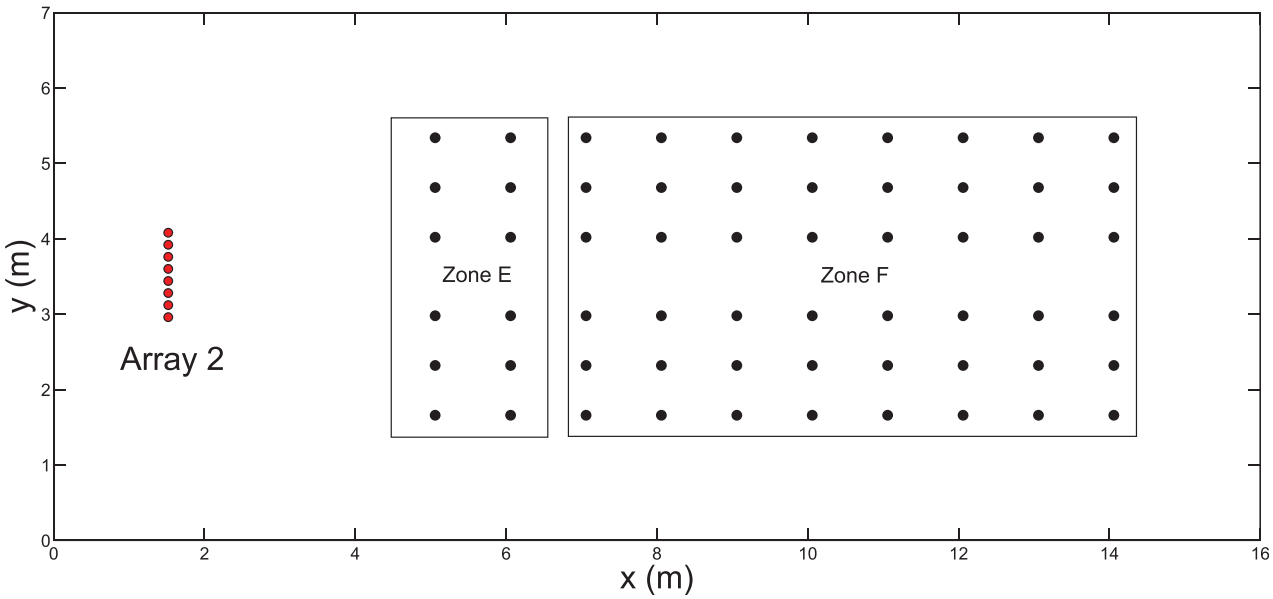

FIG. 20. (Color online) The real-world room setup with the positions of the microphones and the speakers using a single linear array. Two zones E and F were considered with high and low TDOA information taking into account the sensitivity map depicted in Fig. 21.

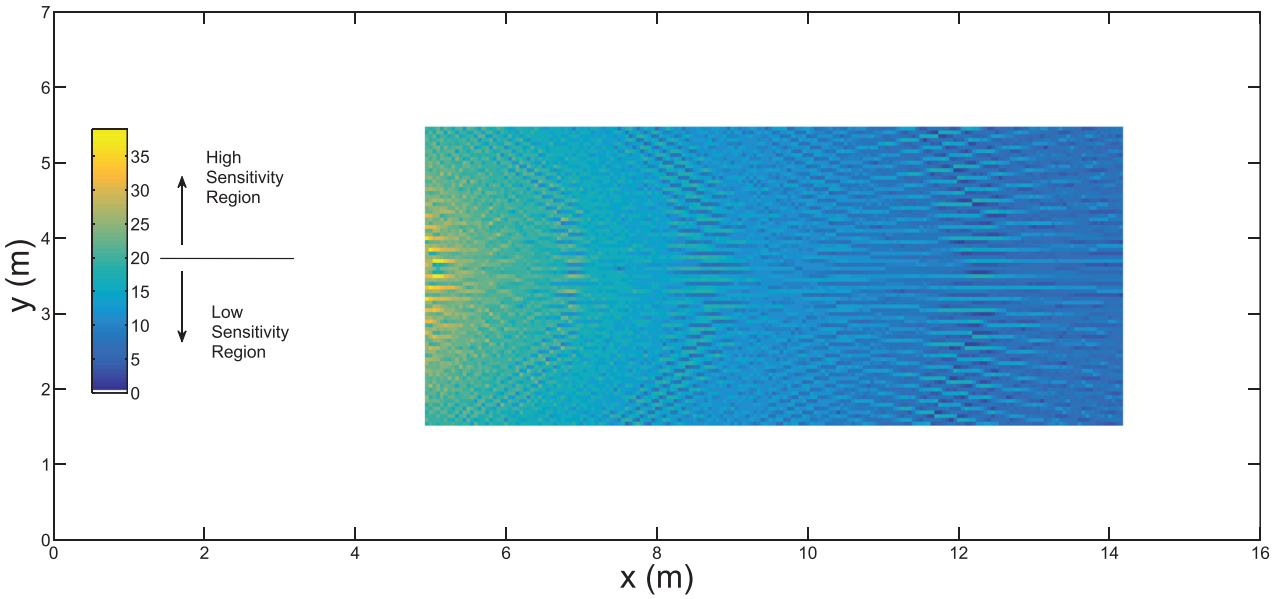

FIG. 21. (Color online) The sensitivity map $\delta(\mathbf{r}_g)$ of the array in Fig. 20 with $\Delta = 0.05$ m and $f_s = 48$ kHz.

midrange valued sensitivity map, as we can see in Fig. 19, and the areas with greater sensitivity are positioned near the arrays 1 and 3. Thus, the performance gap between URG, URG-I and GSG, URG-M, URG-H is also less marked in comparison to the simulated experiments. Specifically, GSG has the best AR in the high sensitivity region, while URG-M and URG-H has a slightly lower overall RMSE.

We also report the localization performance obtained by using a single linear array. The room setup and the power response sensitivity map are depicted in Figs. 20 and 21, respectively. Based on the sensitivity region (Fig. 21), we have selected a high sensitivity region (zone E) and a low sensitivity region (zone F). The grid points calculated with the GSG algorithm cover all the localization area. By comparing Figs. 19 and 21 we can see, from the bar indicating the range of hyperbola intersections, that the use of a single array implies a reduced number of intersections for both the high and low sensitivity regions. Table VI shows the obtained results for the two regions. In this case, the localization totally fails in the low sensitivity region (zone F) for all methods and we can observe a better performance in the high sensitivity region (zone E).

TABLE VI. RMSE (m) and AR (%) (RMSE < 0.2 m) of localization performance for SRP-PHAT with GSG, URG, URG-I, URG-M, and URG-H in a real room with a RT_{60} of 0.9 s using a single linear array.

		GSG	URG	URG-I	URG-M	URG-H
Zone E	RMSE (m)	1.486	2.935	3.740	1.654	1.781
	AR (%)	9.92	8.34	7.22	9.62	9.82
Zone F	RMSE (m)	5.498	4.710	4.385	5.303	5.352
	AR (%)	0.01	0.39	0.57	0.04	0.01

VI. CONCLUSIONS

The paper proposes an algorithm for improved acoustic map computation and spatial search grid design, which leads to an improved SRP-PHAT method. It exploits the geometric properties of the TDOA functions discretization and provides a sensitivity map of the sensor array in use. The advantages of the GSG algorithm for the localization problem of an acoustic source in a reverberant environment are the following.

- It permits the calculation of a sensitivity map, which is a useful tool for identifying the best accuracy zone of a sensor array.
- It allows the design of a spatial grid which is coherent with the acoustic information provided by the sensors array.
- It links all sampling TDOA information from the GCC-PHAT functions into the space resulting in an improved localization in the high sensitivity region.
- SRP-PHAT-GSG performance does not degrade when used with a low spatial resolution grid, due to its spatial resolution scalability properties.
- It permits the reduction of computational cost in those cases in which using the proposed spatial grid is appropriate for the given application or when restricting the search to a high accuracy area for localization.
- It is a useful tool for the reconfiguration of the system, if the setup is not adequate to a specific target.

Experiments were conducted to show the coherent grid design and to analyze the power response sensitivity in the case of a small-array, for different array geometries (linear and randomly distributed sensors), and system parameters: microphone number, sampling frequency, spatial resolution, and microphone distance. Next, by simulations and real-world experimental results, we have shown the importance

of the steered response sensitivity analysis in the localization performance. We have demonstrated that high localization accuracy is achieved in the areas of high sensitivity, while in the low sensitivity region the performance is degraded. Hence, GSG can be used to properly configure the array in order to let the higher sensitivity zones maximally overlap with the target location area.

[1] J. O. Smith and J. S. Abel, "Closed-form least-squares source location estimation from range-difference measurements," IEEE Trans. Acoust. Speech Sign. Process. **35**, 1661–1669 (1987).

[2] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereau, "Real-time passive source localization: A practical linear-correction least-squares approach," IEEE Trans. Speech Audio Process. **9**, 943–956 (2001).

[3] P. Stoica and J. Li, "Source localization from range-difference measurements," IEEE Sign. Process. Mag. **23**, 63–66 (2006).

[4] M. Omologo, P. Svaizer, and R. De Mori, *Spoken Dialogue with Computers* (Academic Press, New York, 1998), Chap. Acoustic Transduction.

[5] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, *Microphone Arrays: Signal Processing Techniques and Applications* (Springer, Berlin, 2001), Chap. Robust Localization in Reverberant Rooms.

[6] P. Aarabi, "The fusion of distributed microphone arrays for sound localization," EURASIP J. Appl. Sign. Process. **4**, 338–347 (2003).

[7] D. B. Ward, E. A. Lehmann, and R. C. Williamson, "Particle filtering algorithms for tracking an acoustic source in a reverberant environment," IEEE Trans. Speech Audio Process. **11**, 826–836 (2003).

[8] P. Pertilä, T. Korhonen, and A. Visa, "Measurement combination for acoustic source localization in a room environment," EURASIP J. Audio Speech Music Process. **2008**, 1–14 (2008).

[9] J. Velasco, D. Pizarro, and J. Macias-Guarasa, "Source localization with acoustic sensor arrays using generative model based fitting with sparse constraints," Sensors **12**, 13781–13812 (2012).

[10] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoust. Speech Sign. Process. **24**, 320–327 (1976).

[11] A. Brutti, M. Omologo, and P. Svaizer, "Multiple source localization based on acoustic map de-emphasis," EURASIP J. Audio Speech Music Process. **2010**, 1–17 (2010).

[12] D. Salvati, C. Drioli, and G. L. Foresti, "Incoherent frequency fusion for broadband steered response power algorithms in noisy environments," IEEE Sign. Process. Lett. **21**, 581–585 (2014).

[13] M. F. Berger and H. F. Silverman, "Microphone array optimization by stochastic region contraction," IEEE Trans. Sign. Process. **39**, 2377–2386 (1991).

[14] D. N. Zotkin and R. Duraiswami, "Accelerated speech source localization via a hierarchical search of steered response power," IEEE Trans. Speech Audio Process. **12**, 499–508 (2004).

[15] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," IEEE Trans. Audio Speech Lang. Process. **15**, 2510–2526 (2007).

[16] L. O. Nunes, W. A. Martins, M. V. S. Lima, L. W. P. Biscainho, M. V. M. Costa, F. M. Gonalves, A. Said, and B. Lee, "A steered-response power algorithm employing hierarchical search for acoustic source localization using microphone arrays," IEEE Trans. Sign. Process. **62**, 5171–5183 (2014).

[17] M. Cobos, A. Marti, and J. J. Lopez, "A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling," IEEE Sign. Process. Lett. **18**, 71–74 (2011).

[18] A. Marti, M. Cobos, J. J. Lopez, and J. Escolano, "A steered response power iterative method for high-accuracy acoustic source localization," J. Acoust. Soc. Am. **134**, 2627–2630 (2013).

[19] M. V. S. Lima, W. A. Martins, L. O. Nunes, L. W. P. Biscainho, T. N. Ferreira, M. V. M. Costa, and B. Lee, "A volumetric SRP with refinement step for sound source localization," IEEE Sign. Process. Lett. **22**, 1098–1102 (2015).

[20] L. O. Nunes, W. A. Martins, M. V. S. Lima, L. W. P. Biscainho, B. Lee, A. Said, and R. W. Schafer, "Discriminability measure for microphone array source localization," in *Proceedings of the International Workshop on Acoustic Signal Enhancement* (2012), pp. 1–4.

[21] M. S. Bartlett, "Smoothing periodograms from time-series with continuous spectra," Nature **161**, 686–687 (1948).

[22] L. Zhang and X. Wu, "On the application of cross correlation function to subsample discrete time delay estimation," Digital Sign. Process. **16**, 682–694 (2006).

[23] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am. **65**, 943–950 (1979).

[24] E. Lehmann and A. Johansson, "Prediction of energy decay in room impulse responses simulated with an image-source model," J. Acoust. Soc. Am. **124**, 269–277 (2008).